Phone-based metric as a predictor for basic personality traits

Mønsted, Bjarke; Mollgaard, Anders; Mathiesen, Joachim

Published in:
Journal of Research in Personality

DOI:
10.1016/j.jrp.2017.12.004

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Journal of Research in Personality 74 (2018) 16–22

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Phone-based metric as a predictor for basic personality traits ☆

Bjarke Mønsted [a,*], Anders Mollgaard [b], Joachim Mathiesen [b]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800, Kgs Lyngby, Denmark
[b] Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 26 September 2017
Revised 14 December 2017
Accepted 28 December 2017
Available online 12 January 2018

Keywords:
Predicting personality
Big Five personality traits
Phone-based metrics
Human dynamics

ABSTRACT

Basic personality traits are believed to be expressed in, and predictable from, smart phone data. We investigate the extent of this predictability using data (n = 636) from the Copenhagen Network Study, which to our knowledge is the most extensive study concerning smartphone usage and personality traits. Based on phone usage patterns, earlier studies have reported surprisingly high predictability of all Big Five personality traits. We predict personality trait tertiles (low, medum, high) from a set of behavioral variables extracted from the data, and find that only extraversion can be predicted significantly better (35.6%) than by a null model. Finally, we show that the higher predictabilities in the literature are likely due to overfitting on small datasets.

1. Introduction

Over the last decades, new data collection methods have provided new opportunities for research on human behavior. Online social networks or personal mobile devices do not only provide real-time data for studies on human activity and interaction, but can also serve as an external validation of e.g. more classical questionnaire- or interview-based studies. For example, the predictability of basic personality traits from smart-phone usage is currently an active area of research Crandall et al. (2010), de Oliveira, Karatzoglou, Concejero Cerezo, Armenta Lopez de Vicuña, and Oliver (2011), LiKamWa, Liu, Lane, and Zhong (2011), Verkasalo, López-Nicolás, Molina-Castillo, and Bouwman (2010), Chittaranjan, Blom, and Gatica-Perez (2011a), Chittaranjan, Blom, and Gatica-Perez (2011b), Williams, Whitaker, and Allen (2012), de Montjoye, Quoidbach, Robic, and Pentland (2013), Sekara and Lehmann (2014), Mollgaard et al. (2016b).

Based on data from the Copenhagen Network Study (CNS), Stopczynski et al. (2014), we use smartphone data to quanitfy the predictability of the Big Five personality traits Digman (1990), openness (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N), commonly called the five factor model and abbreviated as OCEAN. The CNS data is to the best of our knowledge the largest and most detailed study of its kind. Specifically, we use the Big Five Inventory John, Naumann, and Soto (2008), which consists of 44 items. For each item, participants in the CNS study have expressed, on a discrete scale from 1 to 5, how much they agree with a given statement. The personality traits are then computed from a pre-determined linear combination of the 44 answers. Previous research has suggested that smartphone data can be used to predict the Big Five with surprisingly high accuracy de Montjoye et al. (2013). In contrast, we show using a broad range of features extracted from the CNS data that only extraversion can be predicted with some certainty. In the Methods section below and in the appendices, we provide a description of the features (predictor variables) we extract from the smartphone data and further consider their cross-correlations. In the Results section, we use a support vector machine model for the prediction and quantify its relative improvement over a null model where personality scores are randomly assigned. Finally, we briefly compare the scoring system behind the Big Five Inventory against alternative dimensionality reduction techniques in terms of predictability.

2. Methods

We use questionnaire-based data on the personality traits together with phone based-data from 730 freshman students starting in the year 2013 at the Technical University of Denmark. The phone-based data has been collected over a period of 24 months by custom software installed on smartphones given to the participants of the study, Stopczynski et al. (2014). The data consists of telecommunication logs (phone calls, text messages), online social

☆ The study received funding through the University of Copenhagen, UCPH 2016 Excellence Programme for Interdisciplinary Research.
The research reported in this study has not been preregistered in an independent, institutional registry.
* Corresponding author.
E-mail address: bjmo@dtu.dk (B. Mønsted).

https://doi.org/10.1016/j.jrp.2017.12.004
0092-6566/

networks (Facebook connections and interactions), and networks based on physical proximity. The physical proximity is measured through the Bluetooth signal strength, and can be used to monitor face-to-face contacts Sekara and Lehmann (2014). From the GPS data, we obtain information on the geo-spatial mobility Mollgaard, Lehmann, and Mathiesen (2016a). Out of the 730 participants, we only include data from individuals, which, we believe, have used the phone as a primary device. This implies discarding data from users that have written less than 10 text messages, made 5 phone calls or have 100 GPS data points, as well as users with no Facebook friends. These criteria were chosen as a simple heuristic for removing participants who very quickly stopped using the phone, as the subjects remaining after this removal had vastly larger amounts of data. These requirements reduce the number of participants in our study to 636.

For comparison purposes, we consider a list features similar to those in de Montjoye et al. (2013). Furthermore, we repeat our analysis on the part of the *Friends and Family* (FF) dataset, Aharony, Pan, Ip, Khayal, and Pentland (2011), which is publicly available.[1] The FF dataset consists of data from 52 participants, 38 of whom have sufficient call and location data for our analysis according to the selection criteria described above. We finally compare our analysis on both datasets with the results in de Montjoye et al. (2013). Table 1 presents a list of all the features we consider. The feature extraction process is described in detail in the following.

Feature Extraction. The first category of features that we extract consists of basic statistics of calls and texting. For each user, we compute the median and standard deviation of the inter-event time between phone calls, text messages, and combinations thereof. For each of the three interaction forms, we also compute the entropy S_u defined by

$$S_u = \sum_c \frac{n_c}{n_t} \log_2 \frac{n_c}{n_t}, \tag{1}$$

where the index c runs over each unique phone number that the user has contacted, n_c denotes the number of interactions with contact c, and $n_t = \sum_c n_c$ the total number of interactions. The entropy is a general measure of the spread of the interactions. Users with low entropy tend to mainly contact a few individuals while largely ignoring the rest, whereas users with high entropy tend to contact people more equally. We further determine the percentage of a user's calls which were outgoing, as well as the total number of contacts, their ratio to the number of interactions, and the ratio of calls and texts that a user has responded to within an hour of receiving them, and finally the fraction of calls made during the night.

A number of quantities based on location data are also computed. We extract the median and standard deviation of the users' daily distance travelled, their daily radius of gyration (here simplified to be the radius of the smallest circle enclosing all coordinates visited by the user on each day) and the entropy of the time spent in various locations by the user. We identify the locations visited by clustering the GPS points sampled when a user is not moving. A user is defined to not move, if the user's mean speed does not exceed 0.5 m/s in a period between two consecutive GPS points. As the uncertainty on civilian GPS locations can be up to 100 m Zandbergen and Barbeau (2011), a user moving at a speed of 0.5 m/s would need at least 400 s to move a distance larger than two times the uncertainty. For that reason, we consider only GPS points taken even further apart, i.e. 500 s apart. The GPS data points are filtered according to the following procedure. For each user, we include the first recorded GPS data point, we then exclude data points in the subsequent time window of 500 s and then again include the first data point sampled outside this window. From this new data point we repeat the procedure of excluding points in a subsequent window of 500s and so forth. We identify clusters (locations) in the GPS points by use of the DBSCAN algorithm Ester, Kriegel, Sander, and Xu (1996) and we compute the entropy of visits to those clusters by again applying Eq. (1). Finally, we estimate the fraction of time a user spends at home, where home is assumed to be the place where a user spend most of their weeknights.

Another category of features aim to quantify the degree to which a user's behavior follows a temporal pattern. For the call/-text data, we count the number of call/text events for a given user in time bins of 6 h. We then fit an autoregressive series, which best predicts the activity in time bin X_t from previous activities on the form

$$X_t = \mu + \epsilon_t + \sum_{i+1}^{p} \varphi_i X_{t-i}, \tag{2}$$

where μ is the mean activity and ϵ_t is a noise term. These coefficients φ_i are used as features with names like 'AR series coefficient i', where i is the coefficient's lag.

We finally extract a range of features concerning a user's social contacts. This includes their number of Facebook friends and the fraction of the time users spend in the proximity of other participants in the study. This is estimated from repeated automatic scans by the Bluetooth ports. The entropy of the proximity is also calculated similarly to Eq. (1), as well as the time series parameters as described in Eq. (2).

2.1. Classification

We divide the scores on each of the five personality traits into tertiles, i.e. we assign a label of 0, 1 or 2 specifying whether they score low, medium, or high on that trait, corresponding to them lying in the bottom, middle, or upper third, respectively, of all the user scores for that trait. We do this for two reasons - first, this has been done in existing research Chittaranjan et al. (2011a), de Montjoye et al. (2013) and hence allows comparison between our results and those in the literature. Second, although regression approaches have nice accuracy metrics like the mean squared error (MSE), which provides a number for how far from the true values the prediction of the regressor typically is, this measure is not particularly meaningful on ordinal values like personality traits, where e.g. higher extraversion scores mean a person is more extroverted, but there's no precise interpretation for a difference in extraversion score of, say, 0.2.

Our model of choice for predicting the classification labels $\mathbf{Y}$ from the feature vectors $\mathbf{X}$ is a support vector machine (SVM) using a radial basis function (RBF) kernel Hearst, Dumais, Osuna, Platt, and Schölkopf (1998). This model requires that two hyperparameters are fixed - a misclassification cost C and a sharpness γ of the Gaussian basis functions. We take two approaches for feature selection and model fitting, and subsequently compare the results. In both approaches, we use the correlation between phone-metrics and personality traits as a heuristic for feature selection and include the number of features n as a hyperparameter of the model.

In the first approach, we perform a number of cross-validation runs. For each training set introduced during the cross-validation, we first choose the n features with the strongest correlations with the personality traits and perform an extensive grid search in the hyperparameter space. As a consequence, both the hyperparameter values and the features included in the classifier will vary between each cross validation run, potentially making it more difficult to interpret the results. At the same time, however, this ensures that training and test sets are completely separated, and thus that we do not observe overly optimistic results caused by overfitting.

[1] Available at http://realitycommons.media.mit.edu/friendsdataset.html

Table 1
Features included in the classifiers. Table of the features included in the classifiers for each of the Big Five traits abbreviated openness (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N). A dot in a given row/column indicates that the feature corresponding to the row was included in the classifier predicting the personality trait corresponding to the column. Detailed descriptive statistics and visualizations for the features are included in the appendix.

Feature	O	C	E	A	N
Contact entropy using 24-h bins					•
Call duration (median)					•
Call duration (standard deviation)					•
Call inter-event time (standard deviation)	•		•		
Percent of a user's calls initiated by themselves				•	
Call/text contact-interaction ratio				•	
Call/text inter-event time median					•
Call/text inter-event time (std)		•	•	•	•
Ingoing call/text AR series coefficient 13	•				
Ingoing call/text AR series coefficient 4		•			
Number of contacts during the first three months	•		•		
Number of call/text events		•			
Number of texts		•			
Number of Facebook friends	•		•		
Outgoing call/text AR series coefficient 2		•			
Outgoing call/text AR series coefficient 4		•			
Text contact/interaction ratio				•	
Text inter-event time (median)					•
Text inter-event time (standard deviation)		•	•	•	•
Median text response time	•				
Fraction of texts that were outgoing				•	
Fraction of texts responded to within an hour		•		•	•

In a second and less safe approach, following de Montjoye et al. (2013), we use a feature's correlation with a given trait as a heuristic for estimating the importance of the feature. We thus rank the features by their correlations to a given trait, and define another parameter n, denoting the number of features to include, starting with the one most correlated to the personality trait in question. The hyperparameter values and the feature selection are first fixed by performing a grid-search procedure on the full dataset, including into the final classifier the n features with the strongest correlations to the personality trait in the full dataset. This has the disadvantage of being vulnerable to overfitting, especially on smaller datasets, as it allows the classifier to exploit coincidental correlations between phone metrics and personality traits for prediction. On the other hand, this approach has the advantage that hyperparameter values and feature selection is only determined once, which may aid in interpreting results. The values of the hyperparameters C and γ are shown in Table 2, and the features included into the classifiers for the five traits are shown in Table 1.

3. Results

The quality of our classification is measured in terms of the *relative improvement* over a baseline classifier (our null model)

$$S = \frac{f_{\text{classifier}}}{f_{\text{baseline}}} - 1, \tag{3}$$

where f denotes the fraction of correct classifications. The score f_{baseline} is obtained using a null classifier which always predicts the label most frequently occurring in a test set. Using the first approach, outlined above, where hyperparameters are fitted separately on each training sets, we obtain the results shown in Table 3. In general, our relative improvements over the baseline are much lower than those reported in the literature. The only exception is the extraversion trait in the CNS dataset, which at the same time is the only trait that can be predicted significantly better than baseline.

We now compare these results with the less safe approach, where hyperparameters are determined and features selected on the full dataset. For the FF data, we obtain relative improvements of the trait prediction in the range 0.176–0.493 (with a mean improvement of 0.31) based on 10^4 bootstrap samples. For the CNS dataset, we obtain relative improvements over the null model in the range −0.024 to 0.367 (with a mean improvement of 0.11). The results for each trait in each dataset is shown in Table 4.

In de Montjoye et al. (2013) a mean relative improvement of 0.42 is reported, which is significantly above what is reported in another study Chittaranjan, Blom, and Gatica-Perez (2013). We note that significant improvements over a baseline classifier for traits other than extraversion appears contingent on (a) having few data points, and (b) using correlations on the full dataset for feature selection, thus allowing the model to be fit to noise. Hence, it seems likely that earlier reports of high predictability of human

Table 2
Choice of hyperparameters. The values for the hyperparameters C and γ which gave the best prediction in the grid search.

	Openness	Conscient.	Extrav.	Agreeable.	Neuroticism
C	0.8	0.8	1.0	42.0	1.0
γ	0.2	2.0	0.05	0.75	1.0

Table 3
Classifier performance when only correlations internal to the training set are used for variable selection. Comparison of performances on the FF and CNS datasets when the classifier selects features in each cross validation run based only on training set correlations between features and personality traits.

Trait	FF	CNS
Openness	8.3 ± 19.4	3.4 ± 3.4
Conscientiousness	4.8 ± 16.6	-1.2 ± 1.6
Extraversion	-8.9 ± 18.9	35.6 ± 1.3
Agreeableness	-5.2 ± 18.0	-0.0 ± 3.6
Neuroticism	12.6 ± 20.0	-0.8 ± 2.6

Table 4
Performance of the classifier. Comparison of the relative improvement over baseline of our classifier on each of the Big Five traits in our dataset ($n = 636$) with the Friends and Family dataset ($n = 38$).

Trait	FF	CNS
Openness	35.3 ± 11.5	6.2 ± 2.8
Conscientiousness	16.6 ± 13.9	-2.4 ± 2.4
Extraversion	49.3 ± 12.7	36.7 ± 1.1
Agreeableness	17.6 ± 15.1	8.4 ± 3.2
Neuroticism	37.1 ± 17.9	5.5 ± 2.2
Mean	31.2 ± 6.4	10.9 ± 1.1

personality traits from phone metrics have been greatly overestimated due to overfitting enabled by a combination of small sample sizes and a large number of variables. We note that only the extraversion trait appears to be truly predictable from phone-based data. This is in good agreement with common sense, as phones by their nature are devices for inter-human communication. Further, some of the features used in the classifier are expected to be related to extraversion such as the users' number of Facebook friends and the number of new contacts made during the first months of the study.

Based on the Big Five Inventory, the personality traits are computed by reducing the 44 answers to five scores. Any dimensionality reduction of this kind will inevitably lose information available from the full set of answers. We have therefore performed a series of alternative reduction methods on the 44 items to see if we could improve our predictions of the personality traits (see the appendices). Both supervised and unsupervised dimensionality reductions have been used. Among the unsupervised methods, we have tried principal component analysis, independent component analysis and factor analysis. We have applied the methods directly to the answers to the 44 items in order to extract five dimensional objects keeping the most relevant information about the original 44 items. In the unsupervised reduction no information about the features is used. For the supervised reduction method, we try reduce the target variables (the list of items) by finding those items that can be best predicted from the predictor variables (the features). While both the supervised and unsupervised methods improve significantly the quality of our predictions, the overall picture is the same that predominantly items related to extraversion can be predicted with some certainty.

4. Discussion

Using data from the Copenhagen Network Study, which, to our knowledge is the largest dataset simultaneously containing information about the Big Five personality traits and extensive information about smartphone usage patterns, we have shown that the extraversion trait can be predicted significantly better than a null model based on random classification. In contrast, the other personality traits are poorly predicted by our data. Our findings contrast previous studies, which report significant predictabilities across all traits. Given that we have carried out the analysis on datasets of two sizes using two feature selection procedures, and since we obtained high predictabilities only when (a) using full-dataset correlations for variable selection and (b) analyzing a small dataset, the combination of the two appears a likely explanation for the results previously reported in the literature. Regarding the generalizability of our findings, we note that all participants in the study were students at the Technical University of Denmark, and that findings are not necessarily generalizable to the population in general.

5. Availability of data and materials

Data are part of larger study "Social Fabric" involving researchers at the Technical University of Denmark and University of Copenhagen. Due to privacy consideration regarding subjects in our dataset, including European Union regulations and Danish Data Protection Agency rules, we cannot make all data used here publicly available. The data contains detailed information on mobility and daily habits at a high spatio-temporal resolution. We understand and appreciate the need for transparency in research and are ready to make the data available to researchers who meet the criteria for access to confidential data, sign a confidentiality agreement, and agree to work under our supervision in Copenhagen. The "Social Fabric" study was reviewed and approved by the appropriate Danish authority, the Danish Data Protection Agency (Reference number: 2012-41-0664). The Data Protection Agency guarantees that the project abides by Danish law and also considers potential ethical implications. All subjects in the study gave written informed consent.

6. Competing interests

The authors declare that they have no competing interests.

Appendix A. Descriptive statistics of features and target values for the CNS dataset

This section contains descriptive statistics for the applied features and the personality traits. Table A.5 contains key descriptive

Table A.5
Descriptive statistics for the features used for classification. This table summarizes key statistical figures (mean, standard deviation, and min/max values) for the features used. And index is also given to uniquely denote each feature. The indices refer to further graphical information on the features in Fig. A.1. We use the abbreviations *iet* for inter-event time, and *cir* for contact-interaction ratio.

Index	Feature description	μ	σ	min	max
1	Bluetooth daily entropy	4.12	0.32	2.19	4.58
2	Call duration median (s)	51.25	25.11	2.50	353.00
3	Call duration std (s)	284.34	178.72	3.56	1.7e3
4	Call iet std (s)	43e3	15e3	2.4e3	80e3
5	Call percent initiated	0.45	0.06	0.05	0.57
6	Call/text cir	0.03	0.02	2.3e-3	0.31
7	Call/text iet median (s)	512.60	851.12	22.00	11e3
8	Call/text iet std (s)	19e3	11e3	4e3	76e3
9	Incoming activity AR φ_{13}	−2e-3	0.08	−1.59	0.36
10	Incoming activity AR φ_4	0.12	0.07	−0.34	0.64
11	Contacts, first 3 months	53.06	24.93	7	196
12	Number of call/text events	6.2e3	5.6e3	35	41998
13	Number of texts	5e3	5.1e3	20	38,675
14	Number of facebook friends	330.99	182.92	1	1065
15	Outgoing activity AR φ_2	−8.6e-3	0.05	−0.38	0.36
16	Outgoing activity AR φ_4	0.11	0.07	−0.20	0.54
17	Text cir	0.02	0.02	1.2e-3	0.26
18	Text iet median (s)	767.89	3.3e3	19.00	54e3
19	Text iet std (s)	24e3	14e3	3.2e3	88e3
20	Text latency (s)	170.73	123.29	26.00	1.5e3
21	Fraction of outgoing texts	0.45	0.06	0.05	0.57
22	Text response rate	0.34	0.08	0	0.48

figures for the features used in the predictions as listed in Table 1. These include the mean values and standard deviations, as well as the minimum and maximum values for each features. The table also contains a brief description of each feature, as well as an index. These indices can be used to locate a visualization of the distribution of the feature, and information on inter-feature correlations in Fig. A.1. The distribution plots were generated by using a Gaussian kernel density estimation (KDE) procedure to smoothen histograms obtained from the observed features.

Similar details are provided for the big five inventory scores in Table A.6, and the corresponding distributions and correlations are shown in Fig. A.2.

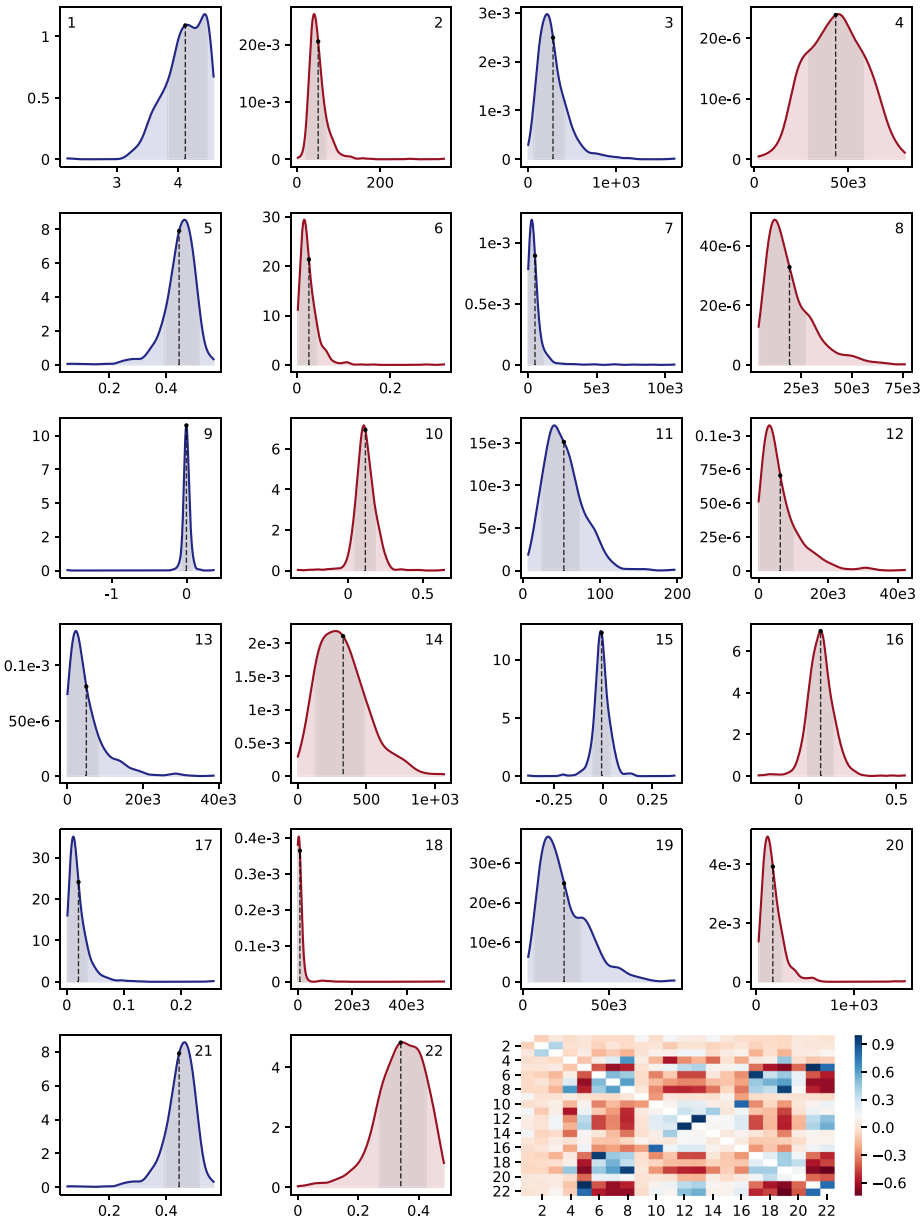

Fig. A.1. Visualisation of distributions and correlations of the observed features. This figure illustrates further details of the distributions of the features listed in Table 1. The black point and vertical line indicate the mean value of the feature, and the darker shaded region correspond to values that are within one standard deviation from the median. The indices in the upper corners of the plots are indices with which the feature names and the exact values of relevant statistical properties can be looked up in Table A.5. In the lower right corners is a heatmap of the pearson correlation coefficients for each pair of features.

Table A.6
Descriptive statistics for the big five inventory scores for the study participants. The tabel shows key statistical figures for the observed scores. The score distributions are shown graphically in Fig. A.2.

Trait	μ	σ	min	max	med
Openness	3.55	0.52	1.90	4.90	3.60
Conscientiousness	3.46	0.56	1.56	4.89	3.44
Extraversion	3.37	0.68	1.50	5.00	3.38
Agreeableness	3.77	0.44	1.89	5.00	3.78
Neuroticism	2.45	0.64	1.12	4.38	2.38

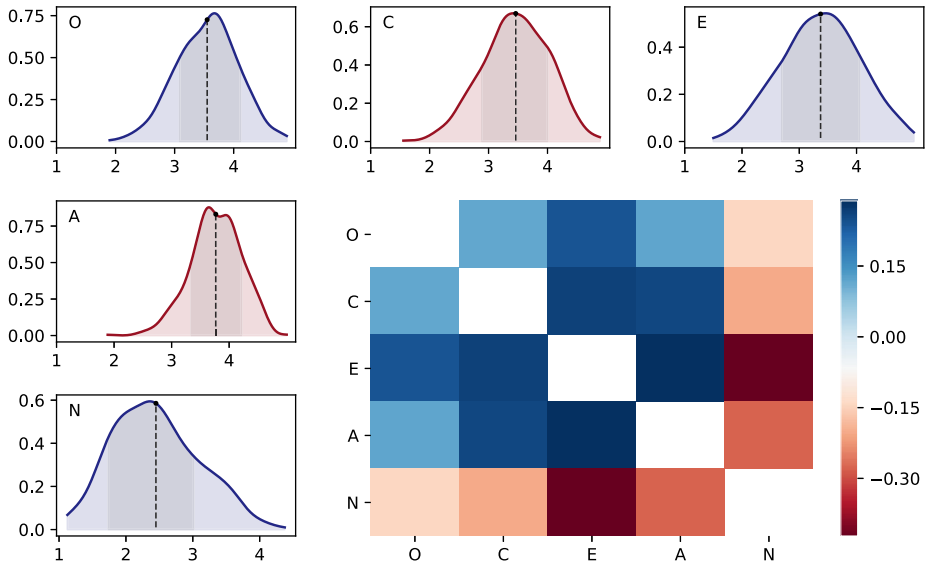

Fig. A.2. Visualization of various properties of the big five inventory scores observed in the study. Visualization of the distributions of scores on each personality trait amongst the study participants, as well as a heat map of inter-trait correlations within the study. The letters in the top-left corners denote the traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism, respectively.

Appendix B. The predictability of alternate linear combinations of questionnaire responses from phone metrics

We investigate the loss of predictability associated with the dimensionality reduction used to compute the Big Five traits from the original 44 questions in the questionnaire, by considering alternative dimensionality reduction techniques. Specifically, we use principal component analysis (PCA), independent component analysis (ICA), factor analysis (FA), and supervised dimensionality reduction (SDR), keeping only the five leading components of each technique. The supervised dimension reduction technique applied here finds the one dimensional projection of the data that has the lowest R^2 value, when training a linear model. The procedure is continued with the additional constraint that the new projections should be orthogonal to all previous projections, such that the result is a low dimensional space specified by an orthogonal basis. The constrained optimization is performed numerically on the training set and then applied to the test set in order to avoid overfitting. See Section C for details on SDR.

Fig. B.3 shows the performance of our classifier in predicting different dimensionality reductions of the 44 questions in the Big Five Inventory. As the figure shows, other dimensionality reduction techniques result in greater personality predictability, indicating that some information related to how people use their phones is contained in their responses to the Big Five questionnaire, but is lost when the Big Five traits are computed from said responses.

To investigate this further, we examined the components of the projection vectors used in each dimensionality reduction technique. In all cases, the projection retaining the greatest predictability was strongly associated with extraversion and in many cases also with neuroticism. For example, Fig. B.4 shows the entries of the ICA vector whose projection had the greatest predictability. Note that the most predictable direction of projection points in a direction corresponding opposite scores of extraversion and neuroticism, consistent with the anticorrelation between the two traits found in the literature Hamburger and Ben-Artzi (2000).

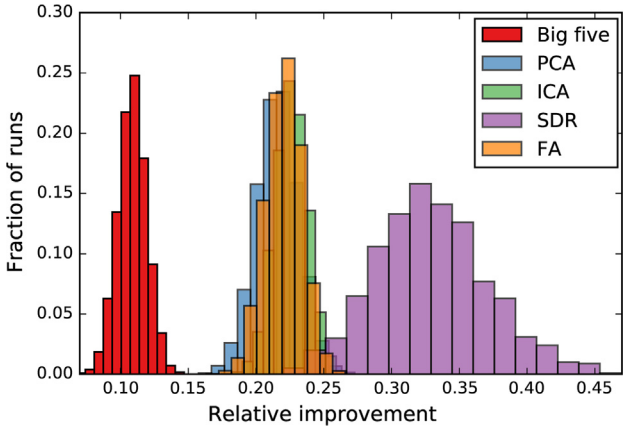

Fig. B.3. Comparison of the predictability of the Big Five personality traits for different dimensionality reduction techniques. For the different reduction techniques, we have averaged predictability score over the five leading components. The plot shows the distribution of relative improvements for 10^4 cross validation runs using various projections of the data. Apart from the Big Five projections (with a mean predictability increase of 0.11 ± 0.01), the unsupervised dimensional reduction techniques include principal component analysis (PCA, 0.22 ± 0.02), independent component analysis (ICA, 0.23 ± 0.01) and factor analysis (FA, 0.22 ± 0.01). The supervised dimensionality reduction (SDR) technique reached an improvement of 0.33 ± 0.04.

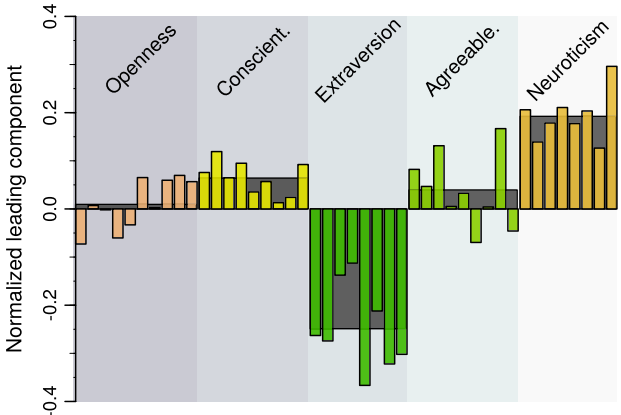

Fig. B.4. The ICA component with the highest predictability. The 44 entries are grouped according to which big five trait the corresponding question is associated with. The wider bars behind show the mean value of each group of questions, thus denoting how strongly associated the ICA component is with each of the five traits.

Appendix C. Supervised dimensionality reduction

In this section we explain in greater detail the supervised dimensionality reduction technique applied in the paper. The goal is to find the projections of the 44 questions, which we can predict the best.

The questionnaire data is represented by a matrix, y_{ij}, where i denotes a person and j denotes a question. Similarly, we have a matrix describing smartphone behavior, x_{ij}, where i denotes a person and j the behavioural variable. The projection vector, p_j, is 44 dimensional and has unit length

$$1 = \sum_j p_j^2. \tag{C.1}$$

It reduces the information in the 44 questions to a single number through an inner product

$$y_i^{(p)} = \sum_j y_{ij} p_j. \tag{C.2}$$

We introduce a linear model to estimate this value based on the behavioural variables

$$y_i^{(p)} = \sum_j x_{ij} \alpha_j + \epsilon_i,$$

where ϵ_i is the error of the model estimate for person i. We aim to train the projection vector, p_j, and the linear model parameters, α_j, such that the coefficient of determination, R^2, is as large as possible. The coefficient is defined as

$$R^2 = 1 - SS_{\text{res}}/SS_{\text{tot}}, \tag{C.3}$$

where

$$SS_{\text{res}} = \sum_i \epsilon_i^2, \tag{C.4}$$

and

$$SS_{\text{tot}} = \sum_i \left(y_i^{(p)} - \bar{y}^{(p)}\right)^2, \tag{C.5}$$

with $\bar{y}^{(p)}$ the average projection over the persons. The training is performed iteratively in two steps. First, we fix the projection vector and optimize for the parameters of the linear model. Then we fix the parameters and optimize for the projection vector. The optimization step is performed using Sequential Least Squares Programming (SLSQP) with the projection vector constrained to unit length. The training converges consistently irrespective of the initialization of the projection vector.

We may then look for the best projection in the 43 dimensional space orthogonal to our first projection. This can either be done by mapping on to these 43 dimensions or simply adding an orthogonality constraint to the optimization. This procedure may be repeated until a satisfying number of projections is obtained.

We have a final note regarding over training. Let us start by counting the number of free parameters in the training step. If the dimension of y is N and the dimension of x is M, then the number of free parameters is $M + N$, since the linear model has an extra parameter for offset, which is canceled by the unit length constraint on the projection vector. For a data set of size S, we need $S \gg M + N$ for proper training. In other words, if too many features of x are included in the SDR scheme, fitting to noise will take place, thereby resulting in worse performance when applying the classifier to a test set. To avoid this over fitting effect, we implement the following procedure to determine the optimal features of x to include. First we partition the data into five test, and training, sets consisting of 80% and 20% of the data, respectively. Within each training set, we find the correlation between the features of x and each of the 44 features of y. For each feature, we compute the product of the p-values corresponding to those correlations, obtaining a value between 0 and 1, where a value of 1 is interpreted as the feature being unrelated to y and lower values indicating stronger associations. We then rank the features according to these values, and keep the n best features for the classification task. We find that $n = 8$ performs the best, since overfitting takes over for larger n, and we therefore use these 8 features for the supervised dimensionality reduction.

References

Aharony, N., Pan, W., Ip, C., Khayal, I., & Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing, 7*, 643–659.

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011a). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing, 17*, 433–450. https://doi.org/10.1007/s00779-011-0490-1 <http://link.springer.com/10.1007/s00779-011-0490-1>.

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011b). Who's who with big-five: analyzing and classifying personality traits with smartphones. In *2011 15th Annual international symposium on wearable computers* (pp. 29–36). IEEE. https://doi.org/10.1109/ISWC.2011.29 <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5959587>.

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing, 17*, 433–450.

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences, 107*, 22436–22441.

de Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. (2013). *Predicting personality using novel mobile phone-based metrics. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7812 LNCS* (pp. 48–55). doi:https://doi.org/10.1007/978-3-642-37210-0_6.

de Oliveira, R., Karatzoglou, A., Concejero Cerezo, P., Armenta Lopez de Vicuña, A., & Oliver, N. (2011). Towards a psychographic user model from mobile phone usage. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (pp. 2191). New York, New York, USA: ACM Press. https://doi.org/10.1145/1979742.1979920 <http://dl.acm.org/citation.cfm?id=1979742.1979920>.

Digman, J. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* <http://www.annualreviews.org/doi/pdf/10.1146/annurev.ps.41.020190.002221>.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* <http://www.aaai.org/Papers/KDD/1996/KDD96-037>.

Hamburger, Y. A., & Ben-Artzi, E. (2000). The relationship between extraversion and neuroticism and the different uses of the internet. *Computers in Human Behavior, 16*, 441–449.

Hearst, M. a., Dumais, S. T., Osuna, E., Platt, J., & Schölkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications, 13*, 18–28. https://doi.org/10.1109/5254.708428 <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=708428>.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of Personality: Theory and Research, 3*, 114–158.

LiKamWa, R., Liu, Y., Lane, N., & Zhong, L. (2011). Can your smartphone infer your mood. In *PhoneSense workshop*, <http://niclane.org/pubs/likamwa_phonesense.pdf>.

Mollgaard, A., Lehmann, S., & Mathiesen, J. (2016a). *General human activity patterns*. Available from arXiv preprint arxiv:1611.08262.

Mollgaard, A., Zettler, I., Dammeyer, J., Jensen, M. H., Lehmann, S., & Mathiesen, J. (2016b). Measure of node similarity in multilayer networks. *PLoS ONE, 11*, 1–10. https://doi.org/10.1371/journal.pone.0157436.

Sekara, V., & Lehmann, S. (2014). The strength of friendship ties in proximity sensor data. *PLoS ONE, 9*, 1–14. https://doi.org/10.1371/journal.pone.0100915. Available from arxiv:1401.5836v3.

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., et al. (2014). Measuring large-scale social networks with high resolution. *PloS one, 9*, e95978. https://doi.org/10.1371/journal.pone.0095978 <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0095978>.

Verkasalo, H., López-Nicolás, C., Molina-Castillo, F. J., & Bouwman, H. (2010). Analysis of users and non-users of smartphone applications. *Telematics and Informatics, 27*, 242–255. https://doi.org/10.1016/j.tele.2009.11.001 <http://www.sciencedirect.com/science/article/pii/S0736585309000793>.

Williams, M. J., Whitaker, R. M., & Allen, S. M. (2012). Measuring individual regularity in human visiting patterns. In *Privacy, Security, Risk and Trust (PASSAT), 2012 international conference on and 2012 international confernece on Social Computing (SocialCom)* (pp. 117–122). IEEE.

Zandbergen, P., & Barbeau, S. (2011). Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. *Journal of Navigation* <http://journals.cambridge.org/abstract_S0373463311000051>.